MYB transcription factors in plants

CATHIE MARTIN (martin@bbsrc.ac.uk)

JAVIER PAZ-ARES (jpazares@samba.cnb.uam.es)

The cloning of the first transcription factor from plants, the C1 gene of maize, indicated that plants use transcription factors that are structurally related to those of animals in their control of gene expression, because C1 showed significant structural homology to the vertebrate cellular proto-oncogene c-MYB. Since 1987, the catalogue of MYB-related transcription factors has increased considerably in size due, primarily, to the ever-expanding number of MYB genes identified in higher plants (Arabidopsis thaliana is estimated to contain more than a hundred MYB genes). In vertebrates, the MYB-related proto-oncogenes comprise a small family with a central role in controlling cellular proliferation and commitment to development. However, while the functions of some plant MYB genes are relatively well understood they are, at present, quite distinct from their animal counterparts.

Most recent reviews on MYB (Box 1) transcription factors have focused on the structure and function of the vertebrate proteins[1–4]. Here we attempt to draw together what is known of the plant MYB transcription factor family, by analysing the structure of plant proteins relative to the prototypic vertebrate proteins, and by summarizing what is known of the function of MYB-related transcription factors in the growth and metabolism of plants. We suggest that the *MYB* family is very important in transcriptional control in higher plants because of the number of genes involved and because of their roles in the control of plant-specific processes.

How similar are plant MYB proteins to their prototypic animal counterparts?

The structural characteristic common to all known MYB proteins is the DNA-binding domain, the signature motif of transcription factor families. MYB proteins containing this domain have been shown to bind to DNA in a sequence-specific manner[5,6]. Additionally, many MYB proteins contain regions with the features of activator domains (generally of the negatively charged type), although for plant MYB proteins it is only in certain cases that there is evidence that these domains serve in transcriptional activation[7–9].

MYB proteins from animals generally contain three repeats (R1, R2 and R3). The MYB DNA-binding domain of plant proteins usually consists of two imperfect repeats of about 50 residues (R2, R3), although exceptionally it can contain only one of these repeats[9] (B. Weisshaar and M. Feldbrugge, pers. commun.). Some MYB proteins identified from fungi also have just two repeats[10,11]. A comparison between the amino acid sequences of representative plant and mammalian MYB proteins reveals that there is greater conservation between the same repeat from different proteins than between the R2 and R3 repeats from the same protein (Fig. 1), in line with the idea that each repeat has a specialized function. Structural analyses of the MYB DNA-binding domain have used c-MYB itself and its vertebrate homologue MYBL2 (also known as B-MYB) most extensively, but, given the high sequence similarity between MYB proteins (Fig. 1), molecular modelling predicts that the structural characteristics of all two-repeat plant MYB proteins are very similar to those of c-MYB (Refs 12, 13). This presumed structural homology has been supported by DNA-binding studies of chimeras of plant MYB proteins and c-MYB, as well as through site-directed mutagenesis of one plant MYB protein, PhMYB3 (Ref. 13).

Each c-MYB repeat folds into a variant of the helix–turn–helix motif, similar to that of the prokaryotic LexA protein, and contains three regularly spaced tryptophan residues (Fig. 1). These tryptophans play a role in the folding of the hydrophobic core of the MYB domain, and are generally conserved in all MYB proteins, although the first tryptophan residue in the R3 repeat is substituted by another aromatic or hydrophobic amino acid in most plant MYB proteins. The solution structure of the complex between the minimal c-MYB DNA-binding domain (R2R3) and its target DNA has revealed a physical interaction between the two repeats. There is a partial overlap in DNA-binding between the repeats, with R2 and R3 positioned towards the 3′ and 5′ parts of the core motif bound by all vertebrate MYB proteins[14]. A probable consequence of this interaction is that it has imposed constraints on repeat co-evolution. In agreement with this, MYB chimaeras prepared by combinations of R2 and R3 repeats from different protein sources generally have reduced binding affinity when compared with their progenitors.

The DNA-binding specificity of plant MYB proteins differs considerably between themselves, as well as from that of the vertebrate MYB proteins[15–19]. For instance, the maize P protein recognizes the motif [C/A]TCC[T/A]ACC similar to that bound by AmMYB305 from *Antirrhinum*, and neither of these proteins appears to bind to the similar vertebrate MYB consensus motif (TAACNG) (Refs 17, 18). PhMYB3 from *Petunia* binds to two sequences, MBSI (TAAC[C/G]GTT) and MBSII (TAACTAAC) (Ref. 18). In the case of PhMYB3, it has been shown that a substitution of a single residue in the R2 recognition helix, Leu44 to Glu (for nomenclature of positions, see Fig. 1), switches the dual DNA-binding specificity to that of c-MYB, and the reciprocal substitution in c-MYB, Glu43 to Leu, gives dual DNA-binding specificity similar to PhMYB3 (Ref. 13). In agreement with experimental data, molecular modelling predicts that the presence in PhMYB3 of Leu, instead of Glu, has two consequences: one direct, due to the change in base-contacting specificity of Leu versus Glu (Ref. 20); and one indirect due to the inability of Leu, in contrast to Glu, to interact electrostatically with a base-contacting residue (Lys40), thereby facilitating the interaction of Lys40 with either of two alternative positions in the target DNA.

Mutations in residues that do not contact bases also affect sequence-specific binding and might account for some of the differences in DNA-binding specificity

Box 1. Glossary of MYB names

Name of MYB protein	Standardized nomenclature	Species
c-MYB (Ref. 52)	HsMYB	*Homo sapiens*
B-MYB/MYBL2 (Ref. 53)	HsMYBL2	
A-MYB/MYBL1 (Ref. 53)	HsMYBL1	
C1 (Ref. 7)	ZmMYBC1	*Zea mays*
P1 (Ref. 17)	ZmMYBP1	
Zm1 (Ref. 39)	ZmMYB1	
Zm38 (Ref. 39)	ZmMYB38	
GAMYB (Ref. 26)	HvMYBGa	*Hordium vulgare*
Am305 (Ref. 22)	AmMYB305	*Antirrhinum majus*
Am340 (Ref. 22)	AmMYB340	
MIXTA (Ref. 24)	AmMYBMx	
GL1 (Ref. 42)	AtMYBGl1	*Arabidopsis thaliana*
AtMYB1 (Ref. 46)	AtMYB1	
AtMYB2 (Ref. 15)	AtMYB2	
MYBPh3 (Ref. 18)	PhMYB3	*Petunia hybrida*
AN2 (Ref. 40)	PhMYBAn2	
MYBf (Ref. 54)	DmMYB	*Drosophila melanogaster*
MYB (Ref. 55)	DdMYB	*Dictyostelium discoides*
CDC5 (Ref. 10)	SpMYBCD5	*Schizosaccharomyces pombe*
FLBD (Ref. 11)	AnMYBFD	*Aspergillus nidulans*

In this review we have used the original gene name when referring to *MYB* genes for which mutations are known. Where MYB genes have been isolated on the basis of sequence homology, we have used a standardized nomenclature giving first the initials of the species, then the term MYB, and then a term describing the particular family member derived from the original descriptions (see references). In the figures illustrating sequence similarities, we use this standardized nomenclature throughout so that MYB proteins from a single species can be readily identified.

between plant MYB proteins[13,20]. Of the eight putative base-contacting residues in MYB proteins, six are fully conserved in all plant MYB proteins, and the remaining two are conserved in at least 80% of these proteins. In particular, the P protein shares all the putative base-contacting residues with c-MYB or PhMYB3 (Fig. 1), but shows a very different DNA-binding behaviour to these two proteins. Therefore, protein context has a significant effect on the specificity properties of base-contacting residues and the strength of their contacts and might also influence the DNA bending or distorting properties of the proteins[21]. In summary, although plant MYB proteins share the homologous MYB domain, differences in their base-contacting residues and in the overall context of their MYB domains produce distinct DNA-binding specificities in different members of the family.

What governs MYB activity in plants?

Gene activity can potentially be regulated at many different stages. Pretranslational control is evident from many differences in the organ-specific and temporal patterns of accumulation of RNA of different plant *MYB* genes[15,22–25] and in response to environmental stimuli, such as light, salt stress or the plant hormones, gibberellic acid and abscisic acid[15,26,27].

Post-translational control can operate through different mechanisms, including cellular redox potential[28], phosphorylation and protein–protein interactions. Thus, the *in vitro* DNA-binding capacity of two plant MYB proteins, AmMYB305 and PhMYB3, has been found to be sensitive to oxidation [R. Solano (1995) PhD Thesis, Univ. de Alcalá], like c-MYB (Ref. 29), but redox control remains to be demonstrated *in vivo* in plants.

Members of the plant MYB family contain several serine or threonine residues, especially in their C-terminal domains, which are possible substrates of kinases, suggesting that phosphorylation can affect the activity of some plant MYB proteins as it does for c-MYB (Refs 4, 22). Phosphorylation might influence DNA binding or transcriptional activation potential. However, experimental evidence for such control in plant MYB proteins is presently lacking, except for AmMYB340, a protein that shows little DNA binding when synthesized *in vivo*, but that recovers high-affinity binding, equivalent to that of the protein synthesized *in vitro*, after treatment of extracts with alkaline phosphatase[29].

MYB proteins can potentially compete for common target motifs. This type of interaction has been suggested to occur in mammalian cells where transcriptional activation by c-MYB (or its retroviral derivative, v-MYB) can be inhibited in some cells by the activity of MYBL2, which can bind to the same DNA sequences[30,31]. Such competition might also occur in plant cells, because MYB proteins with very similar DNA-binding domains can compete for a common target site; this has been shown for two flower-specific MYB proteins, AmMYB305 and AmMYB340 from *Antirrhinum*. The net activation of transcription of their target genes in any particular cell will be dependent on the relative amounts of the two proteins, their relative abilities to bind DNA and their differing abilities to activate transcription[29].

MYB proteins also interact with other transcriptional regulators. Such interactions are widespread for c-MYB, and some are believed to involve interactions with a negative regulatory domain in the C-terminus of the protein that contains a leucine zipper motif. No similar domain has yet been described in any plant MYB protein. However, the C1 protein of maize interacts, in an obligate fashion, with the R protein, or its homologues (B, SN and LC), to promote anthocyanin biosynthesis (see below). R, LC, SN and B contain a bHLH (basic-helix–loop–helix) motif characteristic of MYC transcription factors[32]. Experiments using the yeast two-hybrid system have shown that the C1 protein interacts directly with the B protein, an interaction requiring the N-terminal part of C1 (containing the MYB domain), and the N-terminal domain of B (which does not contain the bHLH domain). B derivatives lacking the bHLH motif still retain their competence to induce anthocyanin biosynthesis and can still bind C1, strongly suggesting that the major role of B is via its interaction with the C1 protein[33].

What do plant MYB transcription factors do?

The three cellular members of the vertebrate MYB family, c-MYB, MYBL1 and MYBL2 recognize similar target motifs and all are believed to play roles in cellular proliferation and are expressed in actively dividing cells before differentiation, albeit in somewhat different tissues[34–36]. In plants there is good evidence for distinct functions for different MYB proteins; some controlling secondary metabolism, some regulating cellular morphogenesis and some serving in the signal transduction pathways responding to plant growth regulators. Within these groups there are subgroups of MYB proteins with overlapping functions as seen with the vertebrate MYB proteins.

Phenylpropanoid metabolism

Phenylpropanoid metabolism is one of the three main types of secondary metabolism in plants involving modification of compounds derived initially from phenylalanine. Through one branch (flavonoid metabolism) it is responsible for the production of a major group of plant pigments (the anthocyanins) and other minor groups (aurones and phlobaphenes) and it also produces compounds that modify pigmentation through chemical interaction with the anthocyanins (co-pigmentation), such as the flavones and flavonols. Flavones and flavonols also serve to absorb ultraviolet light to protect plants. Several flavonoids act as signalling molecules in legumes inducing gene expression in symbiotic bacteria in a species-specific manner, and others act as factors required for pollen maturation and pollen germination in some plant species. A number of flavonoids and related phenylpropanoids (such as stilbenes) also act as defensive agents (phytoallexins) against biotic and abiotic stresses in particular plant species. Another branch of phenylpropanoid metabolism produces the precursors for production of lignin, the strengthening and waterproofing material of plant vascular tissue and one of the principal components of wood. This branch also produces other soluble phenolics, which can serve as signalling molecules, cell-wall crosslinking agents and antioxidants.

(a)

```
HsMYB       LIKGPWTKEEDQRVIELVQKYGPKR.WSVIAKHLKG. RIGKQCRERWHNHLNPE
HsMYBL1     LIKGPWTKEEDQRVIELVQKYGPKR.WSLIAKHLKG. RIGKQCRERWHNHLNPE
HsMYBL2     LVKGPWTKEEDQKVIELVKKYGTKQ.WTLIAKHLKG. RLGKQCRERWHNHLNPE
DdMYB       LVKGAWTKDEDDKVIELVKTYGPKK.WSDIALHLKG. RMGKQCRERWHNHLNPN
DmMYB       LIKGPWTRDEDDMVIKLVRNFGPKK.WTLIARYLNG. RIGKQCRERWHNHLNNP
AnMYBFD     HRRGPWVPEEDQLLLQLVREQGPNNNWVRISQHMHY. RSPKQCRERYHQNLKPS
SpMYBCD5    LKGGAWKNTEDEILKAAVSKYGKNQ.WARISSLLVR. KTPKQCKARWYEWIDPS
AmMYB305    VRKGPWTMEEDLILINYIANHGEGV.WNSLARSAGLK RTGKSCRLRWLNYLRPD
AmMYBMx     VKKGPWTVDEDQKLLAYIEEHGHGS.WRSLPLKAGLQ RCGKSCRLRWANYLRPD
AtMYB1      RVKGPWSKEEDDVLSELVKRLGARN.WSFIARSIPG. RSGKSCRLRWCNQLNPN
AtMYB2      VRKGPWTEEEDAILVNFVSIHGDAR.WNHIARSSGLK RTGKSCRLRWLNYLRPD
AtMYBGl1    YKKGLWTVEEDNILMDYVLNHGTGQ.WNRIVRKTGLK RCGKSCRLRWMNYLSPN
HvMYBGa     LKKGPWTSAEDAILVDYVKKHGEGN.WNAVQKNTGLF RCGKSCRLRWANHLRPN
PhMYB3      LKKGPWTAAEDSILMEYVKKHGEGN.WNAVKRNSGLM RCGKSCRLRWANHLRPN
PhMYBAn2    VRKGAWTEEEDLLLRECIDKYGEGK.WHLVPVRAGLN RCGKSCRLRWLNYLRPH
ZmMYB1      LNRGSWTPQEDMRLIAYIQKHGHTN.WRALPKQAGLL RCGKSCRLRWINYLRPD
ZmMYB38     TNRGAWTKEEDERLVAYIRAHGEGC.WRSLPKAAGLL RCGKSCRLRWINYLRPD
ZmMYBC1     VKRGAWTSKEDDALAAYVKAHGEGK.WREVPQKAGLR RCGKSCRLRWLNYLRPN
ZmMYBPl     LKRGRWTAEEDQLLANYIAEHGEGS.WRSLPKNAGLL RCGKSCRLRWINYLRAD

HsMYB       VKKTSWTEEEDRIIYQAHKRLG  NRWAEIAKLLP..GRTDNAIKNHWNSTMRRK
HsMYBL1     VKKSSWTEEEDRIIYEAHKRLG  NRWAEIAKLLP..GRTDNSIKNHWNSTMRRK
HsMYBL2     VKKSCWTEEEDRIICEAHKVLG  NRWAEIAKMLP..GRTDNAVKNHWNSTIKRK
DdMYB       IKKEAWSDEEDQIIRDQHAIHG  NKWAEIAKFLP..GRTDNAIKNHWNSSMKRV
DmMYB       IKKTAWTEKEDEIIYQAHLELG  NQWAKIAKRLP..GRTDNAIKNHWNSTMRRK
AnMYBFD     LNRDPISAEEGLAIERMVNEMG  RCWAEIARRLG..NRSDNAVKNWWNGNMNRK
SpMYBCD5    IKKTEWSREEDEKLLHLAKLLP  TQWRTIAPIAPIVGRTATQCLERYQKLLDDL
AmMYB305    VRRGNITPEEQLLIMELHAKWG  NRWSKIAKTLP..GRTDNEIKNYWRTRIQKH
AmMYBMx     IKRGPFSLQEEQTIIQLHALLG  NRWSAIASHLP..KRTDNEIKNYWNTHLKKR
AtMYB1      LIRNSFTEVEDQAIIAAHAIHG  NKWAVIAKLLP..GRTDNAIKNHWNSALRRR
AtMYB2      VRRGNITLEEQFMILKLHSLWG  NRWSKIAQYLP..GRTDNEIKNYWRTRVQKQ
AtMYBGl1    VNKGNFTEQEEDLIIRLHKLLG  NRWSLIAKRVP..GRTDNQVKNYWNTHLSKK
HvMYBGa     LKKGAFTPEEERLIIQLHSKMG  NKWARMAAHLP..GRTDNEIKNYWNTRIKRC
PhMYB3      LKKGAFTVEEERIIIELHAKLG  NKWARMAAQLP..GRTDNEIKNYWNTRLKRR
PhMYBAn2    IKRGDFSLDEVDLILRLHKLLG  NRWSLIAGRLP..GRTANDVKNYWNTHLRKK
ZmMYB1      LKRGNFTDEEEEAIIRLHGLLG  NKWSKIAACLP..GRTDNEIKNVWNTHLKKK
ZmMYB38     LKRGNFTADEDDLIVKLHSLLG  NKWSLIAARLP..GRTDNEIKNYWNTHVRRK
ZmMYBC1     IRRGNISYDEEDLIIRLHRLLG  NRWSLIAGRLP..GRTDNEIKNYWNSTLGRR
ZmMYBP1     VKRGNISKEEEDIIIKLHATLG  NRWSLIASHLP..GRTDNEIKNYWNSHLSRQ
```

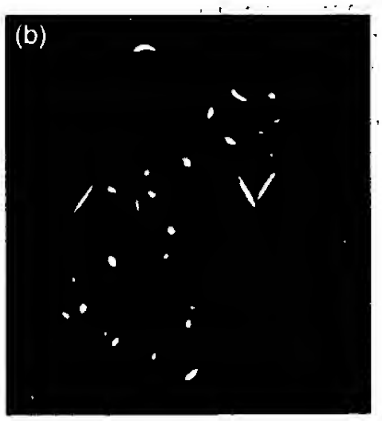

FIGURE 1. The DNA-binding domain of plant MYB proteins: sequence and structure. (a) Sequence alignment of the R2 and R3 repeats of representative MYB proteins from plants, fungi and animals using the CLUSTALV program[51]. To unify nomenclature all MYB proteins have been renamed to add a two letter prefix as a species identifier (Box 1). Residues conserved in all plant MYB proteins are highlighted in white letters. The three regularly spaced tryptophan residues present in each repeat of animal MYB proteins and known to be important in maintaining the hydrophobic core of the DNA-binding domain are labelled with asterisks. The positions corresponding to base-contacting residues of the murine homologue of HsMYB are marked with arrowheads (K39, E43, N47, N90, K93, N94, N97 and S98; where position 1 is the first residue of the R2 repeat) and the size reflects the strength of the contact. The three helices in each repeat[12,14] are shown in the lower part of the figure. The filled box corresponds to the recognition helix. (b) Structure of the MYB–DNA complex. Ribbon plot of the minimal DNA-binding domain of the murine homologue of HsMYB containing the R2 and R3 repeats, binding to its target DNA. The structure represented corresponds to the average of 25 NMR solutions[14]. Molecular modelling[13] predicts similar structures for the different plant MYB proteins represented in Fig. 1(a). The region in the recognition helix of each repeat corresponding to base-contacting residues is highlighted in red.

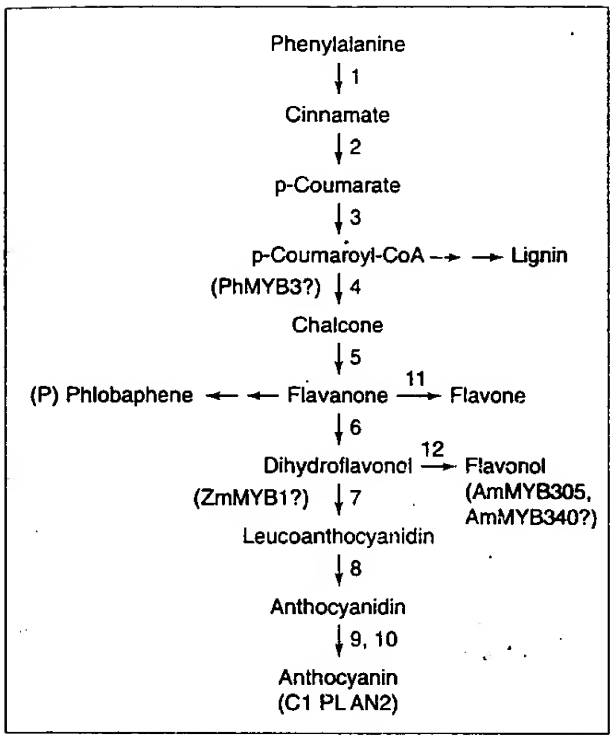

FIGURE 2. Summary of current understanding of the roles of MYB-related transcription factors in controlling phenylpropanoid metabolism from phenyl alanine. MYB-related proteins known to control expression of the subsets of genes encoding enzymes involved in these steps are shown at the end of each pathway. Names with question marks refer to proteins whose role has been demonstrated only biochemically. The functions of MYB proteins without question marks have been demonstrated genetically. P has been shown to control steps 4, 5 and 7, and C1 to control steps 4, 7, 8, 9 and a glutathione-S-transferase (which is involved in transportation of anthocyanins to the vacuole). AN2 controls steps 7, 8 and 10. AmMYB305 and AmMYB340 have been shown to activate steps 1, 5 and 6, PhMYB3 has been shown to activate one gene encoding CHS (step 4) in *Petunia* (CHSJ), and ZmMYB1 has been shown to activate step 7.

MYB proteins are known to play an important role in the control of phenylpropanoid metabolism. The C1 protein activates transcription of genes encoding enzymes involved in the biosynthesis of the anthocyanin pigments in the outer layer of cells of the maize seed endosperm (the aleurone)[7,37,38]. Activation has been demonstrated for five genes in the pathway to anthocyanin (Fig. 2), although C1 probably activates expression of all the genes required specifically for anthocyanin biosynthesis in the aleurone. Activation by C1 involves a partner transcriptional activator in aleurone encoded by the *R* gene[32]. While C1 is active in aleurone, a very similar MYB protein, PL, is functional in controlling anthocyanin biosynthesis in the maize plant (including leaves, stems, and so on) where it interacts with other members of the R-protein family to activate anthocyanin biosynthetic gene expression[23].

In maize, another MYB protein, ZmMYB1 can activate one of the structural genes required for anthocyanin biosynthesis, but not the entire pathway[39], while yet another, ZmMYB38, inhibits C1-mediated activation of the same promoter. It appears that MYB proteins are used to give independent regulation of the structural genes to produce different end products in different cells.

Reiteration of *MYB*-gene function to give metabolic diversity occurs in the control of a branch of flavonoid metabolism producing the red phlobaphene pigments from intermediates in flavonoid metabolism. This pathway is under control of the *P* gene in maize, which encodes a MYB-related protein[17]. The *P* gene product activates a subset of the genes involved in anthocyanin biosynthesis (Fig. 2). The P-binding site is contained within the promoters of these target genes[19], and the *P* gene product does not interact with the R-family proteins and might be able to activate transcription of its target genes alone. So, in maize, at least two different MYB proteins serve to direct flavonoid metabolism along different routes by selective activation of target genes.

In other plant species MYB proteins serve similar roles in the control of phenylpropanoid metabolism as, for example, in *Petunia* flowers where the *AN2* gene product is required for anthocyanin production and has recently been shown to encode a MYB-related product[40] [F. Quattrocchio (1994) PhD Thesis, Vrije Univ. of Amsterdam] but, unlike C1 in maize, AN2 is not required for expression of genes encoding the early steps in anthocyanin biosynthesis (Fig. 2; Ref. 40); perhaps another MYB transcription factor might serve to regulate these early steps. One gene encoding chalcone synthase (CHSJ) can be activated by another MYB protein from *Petunia*, PhMYB3, which is expressed specifically in petal epidermis where anthocyanin pigment is made[18]. Interestingly, there is also strong evidence that AN2 interacts with a bHLH protein to activate transcription[40], although PhMYB3 (unlike C1) can activate transcription alone, suggesting that it does not have an obligate requirement for interaction with a bHLH partner[18].

MYB proteins can also serve to regulate other branches of phenylpropanoid metabolism. In *Antirrhinum majus* and tobacco AmMYB305 (or its orthologue in tobacco) can activate the gene encoding the first enzyme of phenylpropanoid metabolism, phenylalanine ammonia lyase (PAL; Ref. 15). However, the evidence for regulation of other branches of phenylpropanoid metabolism by MYB proteins is, at present, circumstantial and rests on the significance of sequences related to MYB-binding sites in many of the promoters of structural genes in phenylpropanoid metabolism, and lignin biosynthesis[16]. However, some *MYB* genes have been shown to be highly expressed in tissues such as differentiating xylem, supporting the view that they serve roles in controlling the branch of phenylpropanoid metabolism involved in lignin production[41].

Cell shape

The second well-established role for plant *MYB* genes is in the control of cell shape where the *MIXTA* gene of *Antirrhinum* and the orthologous *PhMYB1* gene from *Petunia* have been shown to be essential for developing the conical form of petal epidermal cells (Fig. 3) and the *GL1* gene of *Arabidopsis* has been shown to be essential for the differentiation of hair cells

(trichomes) in some parts of the leaf and in the stem[24,42] [L. Mur (1995) PhD Thesis, Vrije Univ. of Amsterdam]. These two roles might be mechanistically similar because overexpression of *MIXTA* in transgenic tobacco results in trichome formation on petals, suggesting that conical petal cells might be 'trichoblasts' arrested at an early stage in trichome formation (B. Glover and C. Martin, unpublished).

GL1 is required for an initial expansion in the size of the cell that develops into the trichome, and it acts upstream of a number of other genes[43], mutation of which gives rise to cellular outgrowths that do not develop into full, branched trichomes. One, *GL2*, encodes a homeodomain protein that is probably a transcriptional activator of subsequent stages in trichome development[44]. It is, therefore, possible that GL1 is a direct activator of the *GL2* gene. Supporting this idea, the *GL2* gene promoter contains motifs very similar to the binding sites of P and AmMYB305 transcription factors, and these lie in a region shown to affect *GL2* function quantitatively, presumably through affecting the level of *GL2* expression[44]. The conical cells produced by the action of the *MIXTA* gene of *Antirrhinum* resemble the limited outgrowths produced in *Arabidopsis gl2* mutants where trichome formation is aborted. Perhaps the initial stages of trichome formation regulated by GL1 are similar to those regulated by MIXTA. The developmental programme giving rise to conical cells in petals might terminate before the onset of that part of the programme regulated by GL2 in trichome development. In its specification of trichome formation, GL1 is either controlled by or interacts with the product of the *TTG* gene, which is required for trichome formation and anthocyanin production. Overexpression of the maize *R* gene complements the *ttg* mutation leading to the suggestion that the *TTG* gene product is also a R-related protein that interacts with GL1 in a manner analogous to the interaction of C1 and R in maize[45].

Two MYB proteins from fungi, the *CDC5* gene product from *Schizosaccharomyces pombe*[10] and the *FLBD* gene product from *Aspergillus nidulans*[11] can also control aspects of cell shape. The *FLBD* gene product is required for early conidiophore production in *Aspergillus* colonies. The initial branching of the fungal mycelium might have mechanistic similarities to trichome formation, and *FLBD* is thought to activate a cascade of transcription factors for conidiophore production. Clearly, assessment of these similarities in the cellular mode of action of such diverse MYB proteins requires understanding of the specific biochemical processes they activate.

Response to hormones

A more-recently defined role for plant MYB proteins is in hormonal responses during seed development and germination. A barley MYB protein (GAMYB) whose expression is induced by gibberellic acid (GA) has been shown to activate expression of a gene encoding a high pI α-amylase that is synthesized in barley aleurone upon germination for the mobilization of starch in the endosperm[25]. Expression of GAMYB is induced by treatment of aleurone layers with GA and expression of the α-amylase gene is induced subsequently. There is a suggestion that other GA-inducible genes can also respond to activation by MYB proteins during seed germination because MYB-like motifs from other GA-responsive gene promoters have been shown to direct reporter gene expression in response to GA (Ref. 25). There is, as yet, no strong evidence for MYB protein involvement in other GA-induced processes in other parts of the plant although some *MYB* genes are

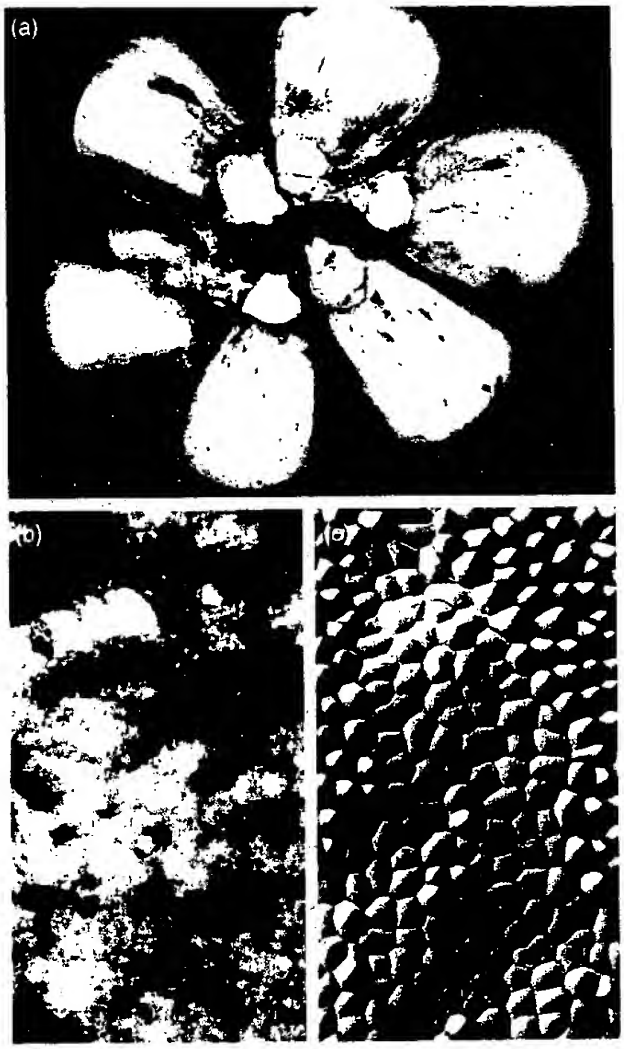

FIGURE 3. Plant phenotypes due to *MYB* gene mutations. (a) Phenotype of unstable allele of the *C1* locus of maize. A transposable element inactivates the gene so the production of anthocyanin pigment in the aleurone layer of the kernel is suppressed. Somatic excision gives rise to pigmented revertant sectors. Photograph courtesy of Brian Scheffler. (b) Phenotype of unstable allele of *MIXTA* locus of *Antirrhinum majus*. Pigmented petal cells are viewed under the microscope. The transposable element suppresses *MIXTA* expression to give paler cells and somatic reversion gives sectors of full colour. The effect of *MIXTA* works not through control of pigment production, however, but through the modification of the optical properties of the petal epidermal cells. (c) The effect of *MIXTA* on cell shape. The *MIXTA* gene controls the formation of conical cells on the petal. Transposon insertion gives rise to flat cells with sectors of revertant conical cells due to somatic excision as revealed by scanning electron microscopy. The conical cells appear more darkly pigmented than the flat cells due to their optimized reflection and refraction properties.

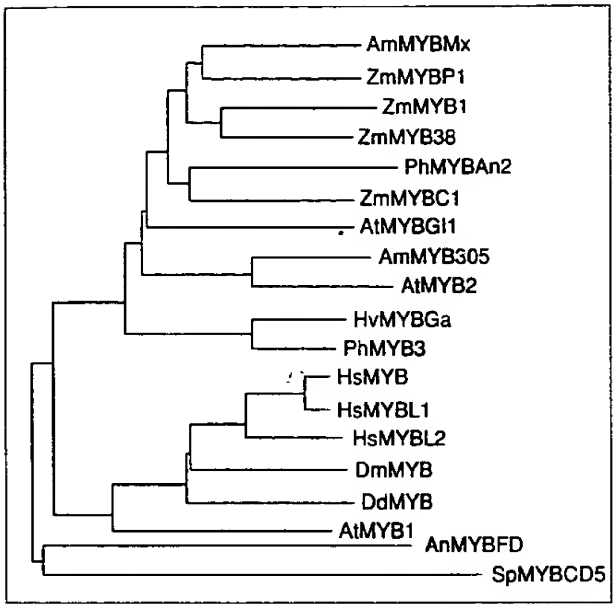

FIGURE 4. Dendrogram of relationships among MYB proteins derived from the matrix of sequence similarities calculated with the CLUSTALV program[51]. Nomenclature for MYB proteins as in Fig. 1.

expressed in response to GA treatment of *Petunia* petals [L. Mur (1995) PhD Thesis, Vrije Univ. of Amsterdam].

Treatment with another plant hormone, abscisic acid (ABA), induces expression of *AtMYB2* in *Arabidopsis*, a *MYB* gene that is also induced in response to dehydration or salt stress[46]. In maize, expression of the *C1* gene is ABA-responsive, where it is involved in the formation of anthocyanin in the developing kernels[27]. AtMYB2 might be responsible for activating expression of some drought-responsive genes because binding by AtMYB2 to the promoter region of a drought- or salt-stress-induced gene, *rd22*, has been demonstrated. The *rd22* gene promoter also contains MYC-recognition sequences suggesting that AtMYB2 can interact with a bHLH protein to induce gene transcription in response to dehydration or salt stress[47].

Cellular proliferation

None of the known functions of MYB proteins in plants bear much similarity to the biological roles of the c-MYB family in vertebrates. Part of the role of c-MYB in promoting cellular proliferation concerns the control of progression of the cell cycle from G1 to S phase through the regulation of CDC2 kinase. c-MYB has been shown to activate transcription from the CDC2 kinase gene promoter in animal cells and so to control the G1–S phase transition, a role that might go part of the way to explain its promotional effects on cellular proliferation[48]. No such role has yet been demonstrated for a plant *MYB* gene product although the *CDC2a* gene from *Arabidopsis* has been shown to contain MYB recognition motifs within its promoter. These sequences lie in regions that enhance the level of expression driven by the promoter as demonstrated by reporter gene fusions[49]. Perhaps there are *MYB* genes in plants with functions more closely analogous to their animal counterparts. There is certainly a subclass of MYB proteins (including AtMYB1 from *Arabidopsis*; Fig. 4) that bear greater structural similarity to vertebrate MYB proteins than to the other plant MYB proteins, and this structural similarity could reflect homologous cellular functions.

Why are there so many plant MYB proteins?

The pervasiveness of *MYB*-related genes in all major groups of eukaryotic organisms suggests that proteins with MYB-like DNA-binding domains developed early to regulate gene expression. Different types of MYB protein might then have evolved as a result of duplication or triplication of the basic repeat unit. It has been proposed that evolution has occurred mostly through modification of regulation of common structural genes, and the separation between different groups of eukaryotes might be accompanied by the differential use of the transcriptional factor classes. This does not, in itself, explain why plants have made such extensive use of MYB proteins, and it might well be that *MYB* genes have been duplicated and their functions expanded in conjunction with the development of novel functions in higher plants.

Although fungi and bryophytes contain MYBs with two repeats, suggesting that R2R3-type proteins were early forms of MYB available for controlling gene expression, it is the size of the R2R3-type *MYB* gene family in higher plants that is particularly remarkable. In one lower plant, the moss *Physcmytrella patens*, the MYB protein family has, in fact, been estimated to be small, with only two to three gene members[50]. *MYB* gene function might have diversified in parallel to increasing complexity in developmental and metabolic pathways as, for example in phenylpropanoid metabolism and also in transcriptional responses to hormones, such as gibberellic acid and abscisic acid, which are, themselves, specialized plant signalling molecules generated from secondary metabolites. So, plants appear to have used R2R3-type MYB transcription factors selectively to control their specialized physiological functions, while in contrast, vertebrates have developed only one small group of MYB proteins to control cellular proliferation and differentiation. However, only about 10% of the plant *MYB* genes have some attributed function, so the full extent of the participation of this transcription factor gene family in plant growth and development is only just becoming realized, as new and diverse functions are defined for its members.

Acknowledgements

We thank the EU Biotechnology programme for funding in this area of research under projects BIOTCT95 10 and BIOZCT93 1010. Thanks also to I. Romero and L. Sanchez-Pulido for help with Figs 1 and 4.

References

1 Graf, T. (1992) *Curr. Biol.* 2, 249–255
2 Thompson, M.A. and Ramsay, R.G. (1995) *BioEssays* 17, 341–350
3 Lyon, J., Robinson, C. and Watson, R. (1994) *Crit. Rev. Oncog.* 5, 373–388

4 Lüscher, B. and Eiseman, R.N. (1990) *Genes Dev.* 4, 2235–2241
5 Howe, K.M., Reakers, C.F.L. and Watson, R.J. (1990) *EMBO J.* 9, 161–169
6 Sakura, H. *et al.* (1989) *Proc. Natl. Acad. Sci. U. S. A.* 86, 5758–5762
7 Paz-Ares, J. *et al.* (1987) *EMBO J.* 6, 3553–3558
8 Goff, S.A., Cone, K.C. and Fromm, M.E. (1991) *Genes Dev.* 5, 298–309
9 Baranowskij, N., Frohberg, C., Prat, S. and Wilmitzer, L. (1994) *EMBO J.* 13, 5383–5392
10 Ohi, R. *et al.* (1994) *EMBO J.* 13, 471–483
11 Wieser, J. and Adams, T.H. (1995) *Genes Dev.* 9, 491–502
12 Frampton, J. *et al.* (1991) *Protein Eng.* 4, 891–901
13 Solano, R. *et al. J. Biol. Chem.* (in press)
14 Ogata, K. *et al.* (1994) *Cell* 79, 639–648
15 Urao, T., Yamaguchi-Shinozaki, K., Urao, S. and Shinozaki, K. (1993) *Plant Cell* 5, 1529–1539
16 Sablowski, R.W.M. *et al.* (1994) *EMBO J.* 13, 128–137
17 Grotewold, E., Drummond, B.J., Bowen, B. and Peterson, T. (1994) *Cell* 76, 543–553
18 Solano, R. *et al.* (1995) *EMBO J.* 14, 1773–1784
19 Li, S.F. and Parish, R.W. (1995) *Plant J.* 8, 963–972
20 Suzuki, M. (1995) *Proc. Jap. Acad. Series B* 71, 27–31
21 Solano, R., Nieto, C. and Paz-Ares, J. (1995) *Plant J.* 8, 673–682
22 Jackson, D. *et al.* (1991) *Plant Cell* 3, 115–125
23 Cone, K.C., Cocciolone, S.M., Burr, F.A. and Burr, B. (1993) *Plant Cell* 5, 1795–1805
24 Noda, K-I., Glover, B.J., Linstead, P. and Martin, C. (1994) *Nature* 369, 661–664
25 Larkin, J.C., Oppenheimer, D.G., Pollock, S. and Marks, M.D. (1993) *Plant Cell* 5, 1739–1748
26 Gubler, F., Kalla, R., Roberts, J.K. and Jacobsen, J.V. (1995) *Plant Cell* 7, 1879–1891
27 Hattari, T. *et al.* (1992) *Genes Dev.* 6, 609–618
28 Myrset, A.H. *et al.* (1993) *EMBO J.* 12, 4625–4633
29 Moyano, E., Martinez-Garcia, J.F. and Martin, C. *Plant Cell* (in press)
30 Foos, G., Grimm, S. and Klempnauer, K-H. (1992) *EMBO J.* 11, 4619–4629
31 Watson, R.J., Robinson, C. and Lam, E.W.F. (1993) *Nucleic Acids Res.* 21, 267–272
32 Ludwig, S.R., Habera, L.F., Dellaporta, S.L. and Wessler, S.R. (1989) *Proc. Natl. Acad. Sci. U. S. A.* 86, 7092–7096
33 Goff, S.A., Cone, K.C. and Chandler, V.L. (1992) *Genes Dev.* 6, 864–875
34 Ausura, M. *et al.* (1992) *Blood* 79, 2708–2716
35 Sitzmann, J., Nobeu-Trauth, K. and Klempnauer, K-H. (1995) *Oncogene* 11, 2273–2279
36 Sitzmann, J., Thrauth, K. and Klempnaeur, K-H. (1995) *J. Cell Biochem.* S19A, 61
37 Paz-Ares, J., Wienand, U., Peterson, P.A. and Saedler, H. (1986) *EMBO J.* 5, 829–833
38 Cone, K.C., Burr, F.A. and Burr, B. (1986) *Proc. Natl. Acad. Sci. U. S. A.* 83, 9631–9635
39 Franken, P. *et al.* (1994) *Plant J.* 6, 21–30
40 Quattrocchio, F. *et al.* (1993) *Plant Cell* 5, 1497–1512
41 Campbell, M.M., Whetton, R.W. and Sederoff, R.R. (1995) *Plant Physiol.* 108 (Suppl.), 28
42 Oppenheimer, D.G. *et al.* (1991) *Cell* 67, 483–493
43 Huilskamp, M., Miséra, S. and Jurgens, G. (1994) *Cell* 76, 555–566
44 Rerie, B., Feldmann, K.A. and Marks, M.D. (1994) *Genes Dev.* 8, 1388–1399
45 Lloyd, A.M., Walbot, V. and Davis, R.W. (1992) *Science* 258, 1773–1775
46 Shinozaki, K., Yamaguchi-Shinozaki, K., Urao, T. and Koizumi, M. (1992) *Plant Mol.* 19, 439–499
47 Iwasaki, T., Yamaguchi-Shinozaki, K. and Shinozaki, K. (1995) *Mol. Gen. Genet.* 247, 391–398
48 Ku, D-H. *et al.* (1993) *J. Biol. Chem.* 268, 2255–2259
49 Chung, S.K. and Parish, R.W. (1995) *FEBS Lett.* 362, 215–219
50 Leech, M.J. *et al.* (1993) *Plant J.* 3, 51–61
51 Higgins, D.G., Bleasby, A.J. and Fuchs, R. (1992) *Comput. Appl. Biosci.* 8, 189–191
52 Majello, B., Kenion, L.C. and Dalla-Favera, R. (1986) *Proc. Natl. Acad. Sci. U. S. A.* 83, 9636–9640
53 Nomura, N. *et al.* (1988) *Nucleic Acids Res.* 16, 11075–11089
54 Katzen, A.L., Kornberg, T.B. and Bishop, J.M. (1985) *Cell* 41, 449–456
55 Stober-Grasser, U. *et al.* (1992) *Oncogene* 7, 589–596

C. Martin *is in the Department of Genetics, John Innes Centre, Norwich Research Park, Colney, Norwich, UK NR4 7UH.*
J. Paz-Ares *is in the CNB-CSIC Campus de Cantoblanco, E-28049 Madrid, Spain.*

Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes

J. L. Riechmann,* J. Heard, G. Martin, L. Reuber, C.-Z. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J. Z. Zhang, D. Ghandehari, B. K. Sherman, G.-L. Yu

The completion of the *Arabidopsis thaliana* genome sequence allows a comparative analysis of transcriptional regulators across the three eukaryotic kingdoms. *Arabidopsis* dedicates over 5% of its genome to code for more than 1500 transcription factors, about 45% of which are from families specific to plants. *Arabidopsis* transcription factors that belong to families common to all eukaryotes do not share significant similarity with those of the other kingdoms beyond the conserved DNA binding domains, many of which have been arranged in combinations specific to each lineage. The genome-wide comparison reveals the evolutionary generation of diversity in the regulation of transcription.

Regulation of gene expression at the level of transcription influences or controls many of the biological processes in a cell or organism, such as progression through the cell cycle, metabolic and physiological balance, and responses to the environment. Development is based on the cellular capacity for differential gene expression and is often controlled by transcription factors acting as switches of regulatory cascades (*1*). In addition, alterations in the expression of genes coding for transcriptional regulators are emerging as a major source of the diversity and change that underlie evolution (*2*).

With the completion of the *Arabidopsis thaliana* genome sequence, the entire complement of genes coding for transcription factors from a plant can be identified and described. Together with the three other eukaryotic genomes that have already been sequenced, it also allows investigation of the similarities and differences in transcriptional regulators among the three eukaryotic kingdoms: plants, animals (*Caenorhabditis elegans* and *Drosophila melanogaster*) (*3, 4*), and fungi (*Saccharomyces cerevisiae*) (*5*). We present such a description and analysis here.

Gene Content and Organization

To characterize the entire complement of transcription factors encoded by the genomes of *Arabidopsis*, *Drosophila*, *C. elegans*, and *S. cerevisiae*, we used a comprehensive list of proteins, domains, and motifs to query the corresponding sequence databases. Transcription factors are usually defined as proteins that show sequence-specific DNA binding and are capable of activating and/or repressing transcription. Although most of the proteins and protein families that were considered in our study fit these criteria, we have also included some other types of transcriptional regulators. Most known transcription factors can be grouped into families according to their DNA binding domain (*6*). Protein domains that are sometimes present in transcription factors, but not necessarily associated with them, have not been included in this genome survey, for example, some zinc coordinating motifs that either are involved in protein-protein interactions or have not yet been functionally characterized.

We searched the *Drosophila*, *C. elegans*, and yeast encoded protein complements (proteomes) using BLAST and motif-finding programs (*7*). Because the complete predicted proteome of *Arabidopsis* was not available at the time of the analysis, we used the entire set of genomic sequences (*7*).

The *Arabidopsis* genome codes for at least 1533 transcriptional regulators, which account for ~5.9% of its estimated total number of genes (Table 1). We identified 635, 669, and 209 transcriptional regulators in the proteomes of *Drosophila*, *C. elegans*, and yeast, respectively (4.5, 3.5, and 3.5%). Thus, the *Arabidopsis* content of transcription factors is 1.3 times that of *Drosophila* and 1.7 times that of *C. elegans* and yeast. These results represent an underestimate of the total number of transcription factors in these organisms. Approximately 40 to 50% of the proteins encoded by each of those genomes cannot be assigned to functional categories on the basis of sequence similarity to proteins of known function (*3, 8–11*). Some of those uncharacterized proteins are expected to be transcriptional regulators (*12, 13*). The large number and diversity of transcription factors in *Drosophila* were proposed to be related to its substantial regulatory complexity (*4*). Applying the same logic to *Arabidopsis* suggests that the regulation of transcription in plants is as complex as that in *Drosophila*. In contrast to *Drosophila* and *C. elegans*, for which a sizable (>25%) fraction of their known transcription factors have been characterized genetically (*14*), only ~5% of those from *Arabidopsis* have been defined by mutation analysis (*15*).

Arabidopsis contains many tandem gene duplications and large-scale duplications on different chromosomes, which might account for >60% of the genome (*9, 10, 16*). Whereas some of these duplications have been followed by rearrangements and divergent evolution, up to 40% of the *Arabidopsis* genes might comprise pairs of highly related sequences (*16*). In that respect, *Arabidopsis* is similar to the three other eukaryotic organisms. The *S. cerevisiae* genome is the result of a complete ancient genome duplication that was followed by extensive gene rearrangements and deletions (*17*). In yeast, ~30% of the genes form duplicate gene pairs. Similarly, duplicated genes account for ~48 and ~40% of the total gene content of *C. elegans* and *Drosophila*, respectively (*11*).

All of the *Arabidopsis* transcription factor gene families are scattered throughout the genome. On average, closely related genes account for ~45% of the total number in the major families (Table 2) (*18*). Gene duplications on different chromosomes are most common (~65%), but duplicated genes are also frequently found at large distances in the same chromosome (~22%) as well as organized in tandem repeats (~13%) (*19*). Clusters of three or more highly related genes are very rare (Table 2).

Transcription Factors Across the Eukaryotic Kingdoms

Two features stand out when comparing the *Arabidopsis* complement of transcriptional regulators with that of the other organisms (Table 3). First, <22% of the *Arabidopsis* transcription factors are zinc-coordinating proteins [belonging to several different families that are thought to have evolved independently (*20*)]. In contrast, zinc-coordinating proteins constitute most of the transcription factors in the three other eukaryotes: ~51% in *Drosophila*, ~64% in *C. elegans*, and 56% in yeast. Second, in *Arabidopsis*, there is no single family of transcription factors that has been so disproportionately am-

Mendel Biotechnology, 21375 Cabot Boulevard, Hayward, CA 94545, USA.

*To whom correspondence should be addressed. E-mail: jriechmann@mendelbio.com

plified as the nuclear hormone receptors in *C. elegans* (~38% of its transcription factors), the C2H2 zinc finger proteins in *Drosophila* (~46%), or the C6 and C2H2 families in yeast (~25% each one). The three largest families of transcription factors in *Arabidopsis*, AP2/EREBP (APETALA2/ethylene responsive element binding protein), MYB-(R1)R2R3, and bHLH (basic helix-loop-helix), each represent only ~9% of the total, and there are several other families with comparable numbers of genes.

Each eukaryotic lineage has its own set of particular transcription factor families and genes [comparing such a small number of genomes represents a limitation for this type of analysis (*21*)] (Table 3). The lineage-specific families are of interest from an evolutionary point of view. According to molecular phylogenetic analyses, plants, animals, and fungi all diverged from a common ancestor during a short period of time, ~1.5 billion years ago (*15*). Thus, it would be expected that most of the transcription factor families would either be shared by the three lineages, if they were present in the common ancestor, or specific to each lineage, if they arose independently following divergence. This is indeed the case (Table 3). Members of lineage-specific families represent 45% of the *Arabidopsis* transcription factors, 47% in *C. elegans*, and 32% in yeast (but only 14% in *Drosophila*, because of its extensive use of the C2H2 zinc finger proteins). Families that are present in all four organisms account for most of the remaining transcription factors in each case.

There are, however, a few exceptions to this expected pattern: some genes and gene families are present in two of the three lineages. Transcription factors and transcription factor families that are present in *Drosophila*, *C. elegans*, and yeast (but are absent from *Arabidopsis*) include the SOX/TCF (SRY-related HMG box/T cell factor) group, the fork head–type/winged-helix proteins, and homologs of the human transcription factor RFX1 (Table 3). The SOX/TCF group, which includes developmental regulators like human SRY (sex-determining region Y) and TCF and the yeast hypoxic-gene regulator ROX1, forms part of the HMG-box (high-mobility group) superfamily of proteins (*22*). In contrast to other HMG-box proteins that act as architectural components of chromatin and have no sequence specificity on their own, the SOX/TCF factors show sequence-specific DNA binding and transactivation activities. There are 14 genes in the *Arabidopsis* genome encoding HMG box–containing proteins, but phylogenetic analyses indicate that none of these proteins belong to the SOX/TCF group (*15*).

In contrast to the examples described above, there does not appear to be any case of transcriptional regulators that are present in both yeast and *Arabidopsis* but absent from animals. This distribution of genes and gene families in the three eukaryotic lineages is in agreement with the notion that animals and fungi are more closely related to each other than to plants (*23*). There are at least three classes of transcription factors that are present in plants and animals but absent from yeast: TUBBY-like (TUB), CPP-like (cysteine-rich polycomb-like protein), and E2F/DP proteins (*13*, *24*, *25*) (Table 3). It remains to be determined whether these classes of genes were specifically lost from the *S. cerevisiae* genome or if they are really absent from the fungal lineage.

There are many transcription factor families that are found only in plants, some of which have been greatly amplified. These include the AP2/EREBP (*26*), NAC (*27*), and WRKY families (*28*); the trihelix DNA binding proteins (*29*); the auxin response factors (ARFs); the Aux/IAA proteins [which do not bind to DNA by themselves, but interact with the ARF proteins (*30*)]; and other smaller families (Table 3). Similarly, animals and yeast have many families of transcription factors that are not found in plants (Table 3).

A lingering question when considering protein families that appear to be exclusive to one lineage is whether their signature domains are true evolutionary innovations or whether their relationships with other proteins have been blurred because their amino acid sequences (but not their three-dimensional structures) have diverged substantially over time. Some of the plant-specific families of transcriptional regulators are characterized by domains that appear to be genuine novelties. For example, the AP2 domain exhibits a new mode of DNA recognition by a β-sheet structure (*31*). Other transcription factors classified as specific to plants, however, might be related to proteins found in other organisms. The plant-specific GRAS proteins might be distant relatives of the animal-spe-

Table 1. Content of transcriptional regulator genes in eukaryotic genomes. The number of genes in each of the eukaryotic genomes is given as an approximate number. This is because the number of genes predicted at the time that a genome is sequenced is always an estimate that is refined over time (*7*).

Organism	Total number of genes	Genes coding for transcriptional regulators	
		Total number	Percentage of total number of genes
A. thaliana	~26,000	1533	5.9
S. cerevisiae	~6,000	209	3.5
C. elegans	~19,000	669	3.5
D. melanogaster	~14,000	635	4.5

Table 2. Gene duplications in *Arabidopsis* transcription factor families. The major families of *Arabidopsis* transcription factors were analyzed for the presence of pairs or groups of highly related genes (*18*). The families analyzed together comprise over 1000 genes. Tandem duplications are arbitrarily defined as those that occur within a sequence distance of 50 kb. If two genes are duplicated in the same chromosome but reside >50 kb apart from each other, they are counted in the "Duplications in the same chromosome" column. (Zn) indicates a zinc coordinating DNA binding motif.

Gene family	Percentage of genes with close homolog	Tandem duplications (%)	Duplications in same chromosome (%)	Duplications in different chromosomes (%)	Number of gene clusters/ number of genes in cluster (chromosome)
MYB-(R1)R2R3	44	7	28	65	0
AP2/EREBP	45	11	39	50	1/3 (4)
bHLH	42	13	13	74	0
NAC	42	27	10	63	1/5 (1)
C2H2 (Zn)	40	9	23	68	1/3 (3)
HB	50	5	24	71	0
MADS	50	30	32	38	1/4 (5)
bZIP	53	13	22	65	1/3 (5)
WRKY (Zn)	33	12	17	71	1/3 (1)
GARP	48	0	8	92	0
Dof (Zn)	37	33	17	50	1/4 (4)
CO-like (Zn)	52	13	13	74	0
GATA (Zn)	50	0	0	100	0
Total	44	13	22	65	NA

cific STATS, based on a similar arrangement of related functional domains (*32*). The trihelix DNA-binding domain, present only in plants, might have evolved from the MYB domain, found in all eukaryotes (*29*).

The two transcription factor families that have been more substantially amplified in *Arabidopsis*, as compared to animals and yeast, are the MYB and the MADS families. The MYB motif consists of a helix-turn-helix structure with three regularly spaced Trp residues. In *Arabidopsis*, almost all of the MYB proteins belong to the MYB-R2R3 class (131 members): they contain two imperfect repeats of the MYB motif (*33*). MYB-R1R2R3 proteins, which are the norm in animals, are rare in *Arabidopsis* (five proteins). The plant-specific R2R3 organization is thought to have evolved from an R1R2R3-type ancestral gene from which the first repeat was lost (*34*). Because the plant MYB-R1R2R3 proteins are more closely related to the animal MYB proteins than to the plant proteins of the R2R3 type, it has been suggested that they might have functions related to those of the MYB proteins in animals, such as the control of cell proliferation (*34, 35*). Conversely, MYB-R2R3 proteins might have evolved to regulate processes specific to plants, including secondary metabolism, responses to plant hormones, and the identity of specific cell types.

In addition to the MYB-(R1)R2R3 proteins, *Arabidopsis* contains additional transcription factors characterized by a more divergent MYB domain, which is present either as a single copy or as a repeat. These proteins form a heterogeneous group and are often referred to as "MYB related." For the purpose of clarity, we have divided the *Arabidopsis* MYB-related proteins into several subclasses in Fig. 1 (*15*).

More distant but also related to the MYB superfamily is a previously unidentified group of proteins that we propose to name "GARP," after maize GOLDEN2, the ARR B-class proteins from *Arabidopsis*, and *Chlamydomonas* Psr1 (*36–39*) (Fig. 1). These proteins appear to be involved in plant-specific processes: GOLDEN2 controls the differentiation of a photosynthetic cell type of the maize leaf, whereas Psr1 is a regulator of phosphorus metabolism.

Arabidopsis also contains many more heat

Table 3. Eukaryotic transcriptional regulators. Number of transcriptional regulators in *Arabidopsis* (*A.t.*), *Drosophila* (*D.m.*), *C. elegans* (*C.e.*), and *S. cerevisiae* (*S.c.*), classified by families on the basis of sequence similarity. The table is nonredundant: proteins are counted only once, regardless of whether they have more than one signature motif. The way in which proteins combine different DNA binding motifs were organized into families is reflected in Fig. 1. Families that are specific to one lineage are indicated in color. Families are listed under "Transcription factors" or "Other transcriptional regulators," as described in the text. However, this distinction is not without problems (for example, the ARID and HMG-box families). Information about the signature motif(s) or sequences that define each family is provided as an InterPro (IPR) or GenBank accession number (*56*). (Zn) indicates a zinc coordinating DNA binding motif. In the bHLH class, only proteins with a discernible basic region were included. "Other" includes some single-copy genes and small families that are not individually mentioned in the text. The results of the database searches (P, motif searches; B, BLAST) and sequence comparisons were inspected by eye. The numbers reported here might therefore differ from other large-scale classifications that are performed automatically (*11*).

Gene family	Predicted # proteins *A.t.*	*D.m.*	*C.e.*	*S.c.*	InterPro or GenBank	Search
Transcription factors						
MYB superfamily						
MYB-(R1)R2R3	**136**	**3**	**2**	**3**	**IPR001005**	**P, B**
MYB-related	**54**	**3**	**1**	**7**	**IPR000818**	**P, B**
AP2/EREBP					IPR001471	B
AP2 subfamily	14	0	0	0		
EREBP subfamily	124	0	0	0		
RAV-like	6	0	0	0		
bHLH	**139**	**46**	**25**	**8**	**IPR001092**	**B**
NAC	109	0	0	0	BAB10725	B
C2H2 (Zn)	**105**	**291**	**139**	**53**	**IPR000822**	**P, B**
HB	**89**	**103**	**84**	**9**	**IPR001356**	**B, P**
MADS	**82**	**2**	**2**	**4**	**IPR002100**	**B**
bZIP	**81**	**21**	**25**	**21**	**IPR001871**	**B**
WRKY (Zn)	72	0	0	0	S72443	B
GARP						
G2-like	41	0	0	0	AAD55941	B
ARR-B class	12	0	0	0	BAA74528	B
C2C2 (Zn)						
Dof	37	0	0	0	CAA66080	B
CO-like	31	0	0	0	AB61139	B
GATA	**28**	**6**	**9**	**10**	**IPR000679**	**B, P**
YABBY	6	0	0	0	AAD38126	B
CCAAT						
HAP2 type	**10**	**1**	**2**	**1**	**A26771**	**B**
HAP3 type	**11**	**2**	**2**	**1**	**P13434**	**B**
HAP4 type	0	0	0	1	S47906	B
HAP5 type	**13**	**3**	**2**	**2**	**Q02516**	**B**
Dr1	**2**	**1**	**1**	**1**	**AAB51375**	**B**
GRAS	12	0	0	0	AAB06418	B
Trihelix	28	0	0	0	S39484	B, P
HSF	**26**	**1**	**1**	**5**	**IPR000232**	**B**
ECP	25	0	0	0	AAC26786	B
ARI	23	0	0	0	AAC49751	B
C3H-type I (Zn)	**17**	**3**	**15**	**3**	**IPR000571**	**P, B**
C3H-type 2 (Zn)	16	0	0	0	CAA65242	B
SBP	16	0	0	0	CAB56581	B
Nin-like	15	0	0	0	CAB61243	B
ARF/AXR1	14	0	0	0	CAA18241	B
TUB	**11**	**2**	**1**	**0**	**IPR000007**	**B**

Gene family	Predicted # proteins *A.t.*	*D.m.*	*C.e.*	*S.c.*	InterPro or GenBank	Search
Transcription factors						
E2F/DP	**8**	**3**	**4**	**0**	**O00716/Q64163**	**B**
CPP (Zn)	**8**	**1**	**1**	**0**	**CAA09028**	**B**
AHL-like	7	0	0	0	AAA20093	B
LIM	**6**	**0**	**0**	**0**	**AAC49750**	**B**
LFY	**1**	**0**	**0**	**0**	AAA32826	B
Other	20	0	0	0		B
NHR (C4) (Zn)	0	21	252	0	IPR001628	B
Ndt-1	0	26	3	0	AAA28325	B
T-BOX	0	8	21	0	IPR001699	B
ETS	0	8	10	0	IPR000418	B
DNL (Zn)	0	4	9	0	IPR001275	B, P
PAIRED w/o HB1	0	5	7	0	IPR001523	B
Runt/CBFα	0	4	1	0	IPR001527	B
NF-kB/Rel/dorsal	0	3	0	0	IPR000451	P, B
Smad	0	3	3	0	BAA76956	B
NF-1/grainyhead	0	2	1	0	CAA33692	B
STAT	0	1	1	0	IPR001217	B
AP-2	0	1	4	0	CAA36842	B
Oct-1/OBF	0	1	1	0	AAA41759	B
ISC-22/Dip3/Bun	0	1	1	0	IPR000580	B
NF-1	0	1	1	0	CAA35853	B
p53	0	1	0	0	CAA42629	B
brinker	0	1	0	0	BAA76710	B
Cut/Zn	0	0	0	37	IPR001138	B, P
Swi1/Swi6	0	0	0	4	CAA45949	B
Copper fist	0	0	0	4	IPR000083	B, P
SP23/MGA2	0	0	0	2	CAA81855	B
ABI3/VP1	0	0	0	2	CAA81951	B
RAP1	0	0	0	1	IPR001357	B
Fork head	**0**	**18**	**15**	**4**	**IPR001766**	**B**
REN	**0**	**1**	**1**	**1**	**NP_002909**	**B**
Other transcriptional regulators						
Novel AA	26	0	0	0	AAC39440	B
HMG-box	**10**	**21**	**15**	**7**	**IPR000910**	**B**
ARID	**4**	**5**	**4**	**2**	**IPR001606**	**B**
JUMONJI	**9**	**2**	**1**	**1**	**T30254**	**B**
PcG: E(z) class	**3**	**1**	**1**	**0**		**B**
PcG: Esc class	**1**	**2**	**1**	**0**	–	**B**
CBF1	0	2	0	0	Q08024	B

shock transcription factors (HSFs) than does *Drosophila*, *C. elegans*, or yeast. Plant HSFs exhibit structural and functional characteristics specific to that lineage (*40*, *41*).

For those transcription factor families that are common to all eukaryotes, how similar are the *Arabidopsis* proteins to those from the other organisms? Each *Arabidopsis* transcription factor was compared to the proteomes of *Drosophila*, *C. elegans*, and yeast by using the BLASTX and BLASTP programs. The analysis revealed that *Arabidopsis* transcription factors do not share significant similarity with those from the other lineages, except in the conserved DNA binding domains that define the respective families. The only *Arabidopsis* proteins that showed similarity beyond the threshold of significance established in the comparison (*42*) were some homologs of the HAP3 subunit of the CCAAT-box binding factor and a MYB-related protein known to be homologous to the *S. cerevisiae* CEF1 and *S. pombe* Cdc5 proteins (*43*, *44*).

Domain Shuffling

The modular nature of transcription factors and the importance of domain shuffling in protein evolution are both well established. The characterization of the entire complement of *Arabidopsis* transcription factors allows consideration of the extent of domain accretion, shuffling, and divergence in these proteins and reveals the relationships among the different families at a genome-wide scale (Fig. 1).

Shuffling of some of the DNA binding domains that are present in all eukaryotes has generated novel transcription factors with plant-specific combinations of modules. This is well illustrated by the homeodomain proteins. In ~50% of the members of the *Arabidopsis* homeobox family, the homeodomain is followed by a leucine zipper (Fig. 1). This combination of motifs is not observed in the yeast or animal homeodomain proteins. The only *Arabidopsis* homeodomain proteins that have an additional motif also found in animal homeodomain proteins are those of the KNOX class, which contain a MEINOX domain (Fig. 1) (*45*). On the other hand, homeodomains in animals are associated with a large variety of motifs, such as the paired and POU-specific domains, the LIM motif, or C2H2 zinc fingers, in combinations that are not present in *Arabidopsis*. Some of these domains (paired and POU) are specific to animals.

Other examples of plant-specific arrangements of common domains include the MADS, YABBY, and ARID families. The ARID (for AT-rich interaction domain) motif is found in animals in a variety of developmental and cell-cycle regulators, like the *Drosophila* Dead ringer and Osa proteins (*46*). In animal ARID proteins, that domain is combined with other motifs, like PHD fingers or the jumonji domain (*47*). In the *Arabidopsis* ARID proteins, the ARID domain is associated with an HMG box, whereas PHD fingers and the jumonji domain form other combinations (Fig. 1). Some animal ARID proteins, like Bright, exhibit sequence-specific DNA binding, whereas others, like Osa, do not. Osa, however, modulates the activity of the SWI/SNF Brahma complex to promote the activation of specific target genes (*46*).

MADS domain proteins in plants were first identified as regulators of floral organ identity and have since been found to control additional developmental processes, such as meristem identity, root development, fruit dehiscence, and flowering time (*48*, *49*). A characteristic of the plant MADS domain proteins that sets them apart from their animal and fungal counterparts is a modular organization containing a distinct coiled-coil domain (K box). The *Arabidopsis* genome sequence, however, has revealed that there is an additional class of plant MADS domain proteins in which the K box is absent (*50*). Phylogenetic analyses indicate that a gene duplication event, ancestral to the divergence of plants and animals, generated two MADS-box gene lineages that are now present in all eukaryotes. In plants, one lineage resulted in MADS proteins with a K box, whereas the other resulted in proteins that lack it (*50*). This conclusion, which was based on sequence phylogeny, is also supported by the structure of the genes. K box–containing MADS-box genes have multiple exons, the

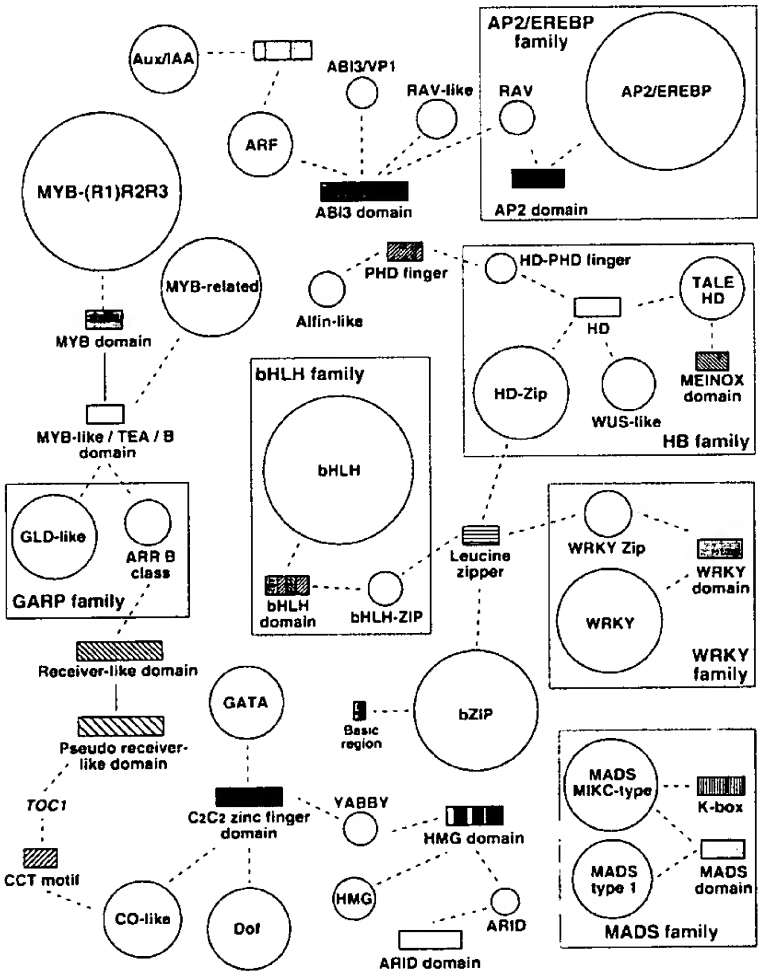

Fig. 1. Relationships and domain shuffling among the different *Arabidopsis* transcription factor families. Gene families are represented by circles, whose size is proportional to the number of members in the family. Domains that have been shuffled and that therefore "connect" different groups of transcription factors are indicated with rectangles, whose size is proportional to the length of the domain. DNA binding domains are colored; other domains (usually protein-protein interaction domains) are shown with hatched patterns. Dashed lines indicate that a given domain is a characteristic of the family or subfamily to which it is connected. Gene names are written in italics. Whereas many of the indicated domain-shuffling events are specific to plants, others likely predate the appearance of the three distinct eukaryotic lineages. For an expanded version of this figure and the information that was used to construct it, see supplemental material (*15*).

MADS box being completely encompassed in one of them. However, analysis of the *Arabidopsis* genomic sequence indicates that MADS-box genes lacking a K box have a simpler structure, with fewer or no introns. *Drosophila* and *C. elegans* each have two MADS-box genes, one per lineage. In *Arabidopsis*, in which at least 82 MADS-box genes can be identified, both classes have been substantially amplified (Fig. 1).

It has been proposed that the complexity in protein domain organization increases with the complexity of the organism (*11*). The above examples of domain shuffling and accretion suggest that, at least among transcription factors, plants are as complex as animals in this respect.

Together with the lineage-specific generation of novel classes of transcription factors or the specific amplification and divergence in one lineage of a common type of regulator, development of novel functions might also result from the organization of transcription factors in novel networks of protein-protein interactions, perhaps as a consequence of domain-shuffling events. For example, the animal-specific Smad proteins depend on interactions with other transcription factors to compensate for their relatively low DNA binding sequence specificity (*51*). These factors include the vertebrate winged-helix protein Fast-1 (winged-helix proteins are found in animals and in fungi) and the *Xenopus* homeodomain proteins Mixer and Milk. The Smad–Mixer/Milk interaction has been proposed to mediate mesoendodermal induction (*52*). All of these Smad-interacting proteins of different classes (Fast1, and Mixer and Milk) share a short Smad-interaction motif (*52*) that appears to be specific to vertebrates: it is not found in *Drosophila*, *C. elegans*, *Arabidopsis*, or yeast proteins. More examples of this kind will be uncovered as the networks of protein-protein interactions among transcription factors are deciphered.

Functional Diversity

The differences in transcription factor content, sequence, and structure among the three eukaryotic lineages are also accompanied by functional diversity. Equivalent or similar biological functions can be controlled by different families of transcription factors in each lineage. Conversely, DNA binding domains that are found in all three eukaryotic kingdoms often control different functions in each one. Developmental regulators illustrate this point. There are also cases, however, in which the involvement of a gene or family in a particular biological function has been maintained across the three lineages (for example, the HSF family).

Pattern formation is an obligate requirement in the development of complex multicellular organisms. In animals, determination of regional identity and specification of the body plan are achieved through the localized activities of homeodomain proteins. Similar functions in plants, meristem patterning and floral organ identity determination, rely on the domain-specific expression of a subset of MADS-box genes (*48*, *49*). Therefore, two different transcription factor families have been used for similar developmental functions in the two lineages.

Patterning depends on a system of axes. The dorsoventral polarity of *Drosophila* has been likened to the dorsoventral asymmetry of zygomorphic flowers and could also be conceptualized as being similar to the adaxial-abaxial polarity of the plant lateral organs. In all of these cases, polarity is established through the regionally localized expression or accumulation of transcription factors, but those belong to different classes. Floral asymmetry in *Antirrhinum* is dependent on the activities of CYC and DICH, two members of the plant-specific family of transcription factors TCP (*53*, *54*). Transcription factors of another plant-specific family, YABBY, are involved in establishing the adaxial-abaxial polarity of the plant lateral organs, together with other genes like PHAN, a MYB-related protein (*55*). In *Drosophila*, embryonic dorsoventral polarity is established through a gradient of Dorsal, a transcription factor of the NF-κB/Rel/Dorsal group (NF-κB, nuclear factor κB). NF-κB/Rel/Dorsal proteins are found in *Drosophila* and mammals but not in *C. elegans*, yeast, or plants.

Conclusion

Each eukaryotic lineage has invented a sizable fraction of its own transcriptional regulators. DNA binding domains that are conserved in sequence and structure have been rearranged in different ways to create novel proteins. The degree of domain shuffling among transcription factors is large. In many instances, families that are common to the three kingdoms have been used for different or novel processes in each of the lineages. The picture that emerges from the comparison of the entire complement of transcription factors of *Arabidopsis*, *Drosophila*, *C. elegans*, and *S. cerevisiae* is one of diversity.

References and Notes

1. M. P. Scott, *Cell* **100**, 27 (2000).
2. S. B. Carroll, *Cell* **101**, 577 (2000).
3. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
4. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
5. A. Goffeau *et al.*, *Nature* **387** (suppl.), 5 (1997).
6. N. M. Luscombe, S. E. Austin, H. M. Berman, J. M. Thornton, review available at http://genomebiology.com/2000/1/1/reviews/001/.
7. The following sequence sets were used: *Drosophila*, 14,080 predicted protein sequences (file aa_gadfly.dros.Z, available at www.fruitfly.org/sequence/download.html); *C. elegans*, 19,101 predicted protein sequences (file WormPep 20, available at www.sanger.ac.uk/Projects/C_elegans/wormpep/); and *S. cerevisiae*, 6308 predicted protein sequences (file orf_trans.fasta.Z, available at ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs/). The complete set of *Arabidopsis* genomic sequences was retrieved from GenBank and analyzed at Mendel Biotechnology. Version 2.0a19MP-WashU of BLAST was used, including the following settings: with BLOSUM62 scoring matrix, with gapping on, without filter, and with other parameters set to default values. Additional information is available as supplemental material (*15*).
8. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
9. X. Lin *et al.*, *Nature* **402**, 761 (1999).
10. K. Mayer *et al.*, *Nature* **402**, 769 (1999).
11. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
12. L. Schauser, A. Roussis, J. Stiller, J. Stougaard, *Nature* **402**, 191 (1999).
13. T. J. Boggon, W.-S. Shan, S. Santagata, S. C. Myers, L. Shapiro, *Science* **286**, 2119 (1999).
14. G. Ruvkun, O. Hobert, *Science* **282**, 2033 (1998).
15. Supplemental material is available at www.sciencemag.org/cgi/content/full/290/5499/2105/DC1.
16. G. Blanc, A. Barakat, R. Guyot, R. Cooke, M. Delseny, *Plant Cell* **12**, 1093 (2000).
17. K. H. Wolfe, D. C. Shields, *Nature* **387**, 708 (1997).
18. The complete set of *Arabidopsis* transcription factors was compared to itself (all against all) with the TBLASTX and BLASTP programs. The BLASTP comparison was used to generate the data summarized in Table 2. A pair of proteins was considered highly similar if they showed >60% amino acid sequence identity along at least two-thirds of the length of one of them.
19. The ordered list of *Arabidopsis* clones that have been used to sequence the genome was obtained from The Arabidopsis Information Resource (TAIR) (www.arabidopsis.org). Those genes that formed related pairs or groups were mapped to the clones, and if they were in the same chromosome, the distance between them was calculated.
20. J. M. Berg, Y. Shi, *Science* **271**, 1081 (1996).
21. E. M. Meyerowitz, *Trends Genet.* **15**, M65 (1999).
22. S. Soullier *et al.*, *J. Mol. Evol.* **48**, 517 (1999).
23. S. L. Baldauf, J. D. Palmer, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11558 (1993).
24. S. Cvitanich *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 8163 (2000).
25. N. Dyson, *Genes Dev.* **12**, 2245 (1998).
26. J. L. Riechmann, E. M. Meyerowitz, *Biol. Chem.* **379**, 633 (1998).
27. M. Aida, T. Ishida, H. Fukaki, H. Fujisawa, M. Tasaka, *Plant Cell* **9**, 841 (1997).
28. T. Eulgem, P. J. Rushton, S. Robatzek, I. E. Somssich, *Trends Plant Sci.* **5**, 199 (2000).
29. Y. Nagano, *Plant Physiol.* **124**, 491 (2000).
30. T. Guilfoyle, G. Hagen, T. Ulmasov, J. Murfett, *Plant Physiol.* **118**, 341 (1998).
31. M. D. Allen, K. Yamasaki, M. Ohme-Takagi, M. Tateno, M. Suzuki, *EMBO J.* **17**, 5484 (1998).
32. D. E. Richards, J. Peng, N. P. Harberd, *Bioessays* **22**, 573 (2000).
33. H. Jin, C. Martin, *Plant Mol. Biol.* **41**, 577 (1999).
34. E. L. Braun, E. Grotewold, *Plant Physiol.* **121**, 21 (1999).
35. H. Kranz, K. Scholz, B. Weisshaar, *Plant J.* **21**, 231 (2000).
36. L. N. Hall, L. Rossini, L. Cribb, J. A. Langdale, *Plant Cell* **10**, 925 (1998).
37. S. Makino *et al.*, *Plant Cell Physiol.* **41**, 791 (2000).
38. D. D. Wykoff, A. R. Grossman, D. P. Weeks, H. Usuda, K. Shimogawara, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 15336 (1999).
39. The similarity among these proteins was probably not realized before because the sequence of the published maize Golden2 is not available from GenBank.
40. E. Czarnecka-Verner, C.-X. Yuan, K.-D. Scharf, G. Englich, W. B. Gurley, *Plant Mol. Biol.* **43**, 459 (2000).
41. F. Schöffl, R. Prändl, A. Reindl, *Plant Physiol.* **117**, 1135 (1998).
42. Each *Arabidopsis* transcription factor was compared by BLASTX and/or BLASTP to a pooled data set that combined the proteomes of *Drosophila*, *C. elegans*, and yeast. A default threshold of $P < 10^{-15}$ was established for the comparison. HSPs with a *P* value

below that threshold were inspected by eye. To be considered significantly similar, the two proteins had to show >50% identity over a region of at least 75% of the length of one of them.
43. T. Hirayama, K. Shinozaki, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13371 (1996).
44. C. G. Burns, R. Ohi, A. R. Krainer, K. L. Gould, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 13789 (1999).
45. T. R. Bürglin, *Dev. Genes Evol.* **208**, 113 (1998).
46. R. D. Kortschak, P. W. Tucker, R. Saint, *Trends Biochem. Sci.* **25**, 294 (2000).
47. D. Balciunas, H. Ronne, *Trends Biochem. Sci.* **25**, 274 (2000).
48. J. L. Riechmann, E. M. Meyerowitz, *Biol. Chem.* **378**, 1079 (1997).
49. G. Theissen *et al.*, *Plant Mol. Biol.* **42**, 115 (2000).
50. E. R. Alvarez-Buylla *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5328 (2000).
51. Y. Shi *et al.*, *Cell* **94**, 585 (1998).
52. S. Germain, M. Howell, G. M. Esslemont, C. S. Hill, *Genes Dev.* **14**, 435 (2000).
53. D. Luo *et al.*, *Cell* **99**, 367 (1999).
54. P. Cubas, N. Lauter, J. Doebley, E. Coen, *Plant J.* **18**, 215 (1999).
55. J. Bowman, *Curr. Opin. Plant Biol.* **3**, 17 (2000).
56. InterPro (www.ebi.ac.uk/interpro/) is a database that integrates protein domain and motif sequence patterns from other databases, like PROSITE, Pfam, and PRINTS.
57. We acknowledge the work of all those who have participated in the Arabidopsis Genome Initiative (AGI), as well as the AGI policy of immediate release of sequence data, which made possible the analysis presented here. We thank all of our colleagues at Mendel Biotechnology for their input and work in our functional genomics research program and E. Meyerowitz and F. Ausubel for discussions and comments on the manuscript.

19 October 2000; accepted 14 November 2000

Orchestrated Transcription of Key Pathways in *Arabidopsis* by the Circadian Clock

Stacey L. Harmer,[1] John B. Hogenesch,[2] Marty Straume,[3] Hur-Song Chang,[4] Bin Han,[4] Tong Zhu,[4] Xun Wang,[4] Joel A. Kreps,[4] Steve A. Kay[1,2]*

Like most organisms, plants have endogenous biological clocks that coordinate internal events with the external environment. We used high-density oligonucleotide microarrays to examine gene expression in *Arabidopsis* and found that 6% of the more than 8000 genes on the array exhibited circadian changes in steady-state messenger RNA levels. Clusters of circadian-regulated genes were found in pathways involved in plant responses to light and other key metabolic pathways. Computational analysis of cycling genes allowed the identification of a highly conserved promoter motif that we found to be required for circadian control of gene expression. Our study presents a comprehensive view of the temporal compartmentalization of physiological pathways by the circadian clock in a eukaryote.

Circadian rhythms control processes ranging from human sleep-wake cycles to cyanobacterial cell division. This is made possible by the circadian clock, an internal biochemical oscillator. The circadian clock allows organisms to anticipate daily changes in the environment such as the onset of dawn and dusk, providing them with an adaptive advantage (*1*). Physiological processes regulated by the clock in higher plants include photoperiodic induction of flowering (*2*) and rhythmic hypocotyl elongation, cotyledon movement, and stomatal opening (*3*). A small number of genes regulated by the clock have been found in an essentially serendipitous fashion (*4*, *5*). However, a global examination of genes controlled by the clock in plants, or in any eukaryote, has been lacking.

[1]Department of Cell Biology, Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA. [2]Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, La Jolla, CA 92121, USA. [3]Center for Biomathematical Technology, NSF Center for Biological Timing, Division of Endocrinology and Metabolism, Department of Internal Medicine, University of Virginia, Charlottesville, VA 22904, USA. [4]Novartis Agricultural Discovery Institute, 3115 Merryfield Row, San Diego, CA 92121, USA.

*To whom correspondence should be addressed. E-mail: stevek@scripps.edu

The circadian clock regulates hundreds of genes. We have used highly reproducible oligonucleotide-based arrays (*6*) to determine steady-state mRNA levels in *Arabidopsis* at 4-hour intervals during the subjective day and night. We examined temporal patterns of gene expression in *Arabidopsis* plants under constant light conditions using GeneChip arrays representing about 8200 different genes. We hybridized duplicate microarrays with biotin-labeled probes derived from plant tissues harvested every 4 hours over 2 days (*7*). Reproducibility between arrays was excellent (Web fig. 1) (*8*). The mean hybridization signal strength and the standard error of the mean for each probe set at each time point were calculated from the duplicate hybridizations.

To objectively determine which genes exhibited a circadian pattern of expression, we empirically tested for statistically significant cross-correlation between the temporal expression profiles of each probe set and cosine waves of defined period and phase. Genes with a greater than 95% probable correlation with a cosine test wave with a period between 20 and 28 hours were scored as circadian-regulated (*9*). This analysis is independent of signal strength and imposes no minimal change in amplitude. According to this criterion, 494 probe sets, representing 453 genes or 6% of the genes on the chip, were classified as cycling (Web table 1) (*8*); 28% of these genes have not been characterized, and no conclusions can be drawn about their function. More than 20 of the known genes we found to be clock-regulated have been previously reported to be under circadian control (*3*, *10*), validating our experimental methods.

We placed the cycling genes into phase clusters of peak expression time. All six possible phases (given our 4-hour time resolution) were well represented, although there were fewer genes peaking at CT16 (*11*) than in other phases [Web table 1 and Web fig. 2 (*8*)]. This is in contrast to cyanobacteria, in which 80% of circadian-regulated genes peak near subjective dusk (*12*). Many of the genes we found to cycle can be clustered into functional groups on the basis of their known and predicted physiological roles.

Clock-controlled anticipation of dawn and dusk. A large cluster of genes implicated in the light-harvesting reactions of photosynthesis were found to be under clock control. mRNAs encoding four LHCA and seven LHCB proteins, chlorophyll binding proteins that funnel light energy to the reaction centers of photosystems I and II, were cycling (Fig. 1A). Also, mRNA encoding an enzyme (protoporphyrin IX magnesium chelatase) involved in the synthesis of their ligand, chlorophyll, was cycling (Web table 1) (*8*). Seven photosystem I reaction center genes and three photosystem II reaction center genes were likewise cycling (Fig. 1B). These 22 photosynthesis genes exhibit striking coregulation, with most peaking around midday at CT4 (*9*). Two *LHC* genes, the reaction center gene *PSAD1*, and the magnesium chelatase gene have been previously reported to cycle (*10*, *13*).

Light also regulates growth and development and resets the circadian clock. Genes encoding phytochrome B (*PHYB*), cryptochrome 1 (*CRY1*), cryptochrome 2 (*CRY2*), and phototropin (*NPH1*) (Web fig. 3A) (*8*) were clock-regulated. Homologs of the blue light photoreceptor genes *CRY1* and *CRY2* are also clock-controlled in animals (*14*). Downstream mediators of phototransduction pathways, *SPA1* and *RPT2*, were also clock-